Novelty Assessment Report

Paper: On The Fragility of Benchmark Contamination Detection in Reasoning Models

PDF URL: https://openreview.net/pdf?id=bhR00j6Mku

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Leaderboards for large reasoning models (LRMs) have turned evaluation into a competition, incentivizing developers to optimize directly on benchmark suites. A shortcut to achieving higher rankings is to incorporate evaluation benchmarks into the training data, thereby yielding inflated performance, known as benchmark contamination. Despite that numerous contamination detection approaches have been proposed, surprisingly, our studies find that evading contamination detections for LRMs is alarmingly easy. We focus on the two scenarios where contamination may occur in practice: (I) when the base model evolves into LRM via supervised fine-tuning (SFT) and reinforcement learning (RL), we find that contamination during SFT can be originally identified by contamination detection methods. Yet, even a brief Group Relative Policy Optimization (GRPO) training can markedly \textbf{conceal contamination signals} that most detection methods rely on. Further empirical experiments and theoretical analysis indicate that Proximal Policy Optimization (PPO) style importance sampling and clipping objectives are the root cause of this detection concealment, indicating that \textbf{a broad class of RL methods} may inherently exhibit similar concealment capability; (II) when SFT contamination with CoT is applied to advanced LRMs as the final stage, most contamination detection methods \textbf{perform near random guesses}. Without exposure to non-members, contaminated LRMs would still have more confidence when responding to those unseen samples that share similar distributions to the training set, and thus, evade existing memorization-based detection methods. Together, our findings reveal the unique vulnerability of LRMs evaluations: Model developers could easily contaminate LRMs to achieve inflated leaderboards performance while leaving minimal traces of contamination, thereby strongly undermining the fairness of evaluation and threatening the integrity of public leaderboards. This underscores the urgent need for advanced contamination detection methods and trustworthy evaluation protocols tailored to LRMs.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Benchmark Contamination Detection in Large Reasoning Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **24 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Contamination Detection Methods**
- **Contamination Evasion and Detection Robustness**
- **Contamination-Resistant Benchmark Design**
- **Domain-Specific Contamination Studies**
- **Contamination Impact and Mitigation Strategies**
- **Contamination Surveys and Empirical Analyses**
- **Evaluation Methodology and Benchmark Validity**

Complete Taxonomy Tree

- Benchmark Contamination Detection in Large Reasoning Models Survey Taxonomy
- Contamination Detection Methods
  - Black-Box Detection Approaches
  - Statistical and Probabilistic Detection (3 papers)
    - [1] Detecting pretraining data from large language models (Shi, 2023) View paper
    - [21] Proving Test Set Contamination in Black Box Language Models (Meister, 2023) View paper
    - [28] Estimating contamination via perplexity: Quantifying memorisation in language model evaluation (Li Yu-Cheng, 2023) View paper
  - Performance-Based and Behavioral Detection (4 papers)
    - [3] Data contamination quiz: A tool to detect and estimate contamination in large language models (Golchin, 2025) View paper
    - [7] ConStat: Performance-Based Contamination Detection in Large Language Models (MÃ¼ller, 2024) View paper
    - [14] PaCoST: Paired Confidence Significance Testing for Benchmark Contamination Detection in Large Language Models (Zhang Hui-xuan, 2024) View paper
    - [35] Simulating Training Data Leakage in Multiple-Choice Benchmarks for LLM Evaluation (Naila Shafirni Hidayat, 2025) View paper
  - Interactive and Dynamic Evaluation Detection (2 papers)
    - [36] Treeeval: Benchmark-free evaluation of large language models through tree planning (Li Xiang, 2025) View paper
    - [39] KIEval: A Knowledge-grounded Interactive Evaluation Framework for Large Language Models (Yu, 2024) View paper
  - White-Box Detection Approaches
  - Training Data Analysis and Retrieval (2 papers)
    - [19] Investigating Data Contamination in Modern Benchmarks for Large Language Models (Deng Chun-yuan, 2023) View paper
    - [42] Training data leakage analysis in language models (Inan, 2021) View paper

  - Embedding and Kernel-Based Detection (1 papers)
    - [25] How Contaminated Is Your Benchmark? Quantifying Dataset Leakage in Large Language Models with Kernel Divergence (Choi, 2025) View paper
  - Fine-Tuning and RL-Stage Detection (2 papers)
    - [22] Detecting Data Contamination from Reinforcement Learning Post-training for Large Language Models (Wang Tian, 2025) View paper
    - [45] DICE: Detecting In-distribution Contamination in LLM's Fine-tuning Phase for Math Reasoning (Tu, 2024) View paper
- Contamination Evasion and Detection Robustness
  - Evasion Techniques and Vulnerabilities ★ (2 papers)
  - [0] On The Fragility of Benchmark Contamination Detection in Reasoning Models (Anon et al., 2026) View paper
  - [16] Evading Data Contamination Detection for Language Models is (too) Easy (MÃ¼ller, 2024) View paper
  - Detection Method Evaluation and Limitations (2 papers)
  - [12] Towards data contamination detection for modern large language models: Limitations, inconsistencies, and oracle challenges (Samuel Vinay, 2025) View paper
  - [31] The Emperor's New Clothes in Benchmarking? A Rigorous Examination of Mitigation Strategies for LLM Benchmark Data Contamination (Sun, 2025) View paper
- Contamination-Resistant Benchmark Design
  - Dynamic and Continuously Updated Benchmarks (5 papers)
  - [2] LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code (Jain, 2024) View paper
  - [20] AntiLeakBench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge (Wu Xiaobao, 2025) View paper
  - [23] BeyondBench: Benchmark-Free Evaluation of Reasoning in Language Models (Srivastava Gaurav, 2025) View paper
  - [24] Dynamic Benchmarking of Reasoning Capabilities in Code Large Language Models Under Data Contamination (Chen Si-min, 2025) View paper
  - [27] Is Your Benchmark (Still) Useful? Dynamic Benchmarking for Code Language Models (Wu Xiao, 2025) View paper
  - Private and Secure Evaluation Protocols (1 papers)
  - [26] Truce: Private benchmarking to prevent contamination and improve comparative evaluation of llms (Chandran, 2024) View paper
  - Contamination-Free Benchmark Construction (3 papers)
  - [29] MMLU-CF: A Contamination-free Multi-task Language Understanding Benchmark (Qihao Zhao, 2024) View paper
  - [33] Putnam-AXIOM: A Functional and Static Benchmark for Measuring Higher Level Mathematical Reasoning in LLMs (Miranda, 2025) View paper
  - [37] RealMath: A Continuous Benchmark for Evaluating Language Models on Research-Level Mathematics (Zhang Jie, 2025) View paper
  - Benchmark Variation and Perturbation (3 papers)
  - [13] Rethinking Benchmark and Contamination for Language Models with Rephrased Samples (Yang Shuo, 2023) View paper
  - [48] LastingBench: Defend Benchmarks Against Knowledge Leakage (Fang Yixiong, 2025) View paper
  - [49] VAR-MATH: Probing True Mathematical Reasoning in LLMS via Symbolic Multi-Instance Benchmarks (Yao Jian, 2025) View paper
- Domain-Specific Contamination Studies
  - Code and Programming Benchmarks (1 papers)
  - [30] Beyond Correctness: Benchmarking Multi-dimensional Code Generation for Large Language Models (Zheng Jiasheng, 2024) View paper
  - Mathematical and Scientific Reasoning (2 papers)
  - [17] PHYBench: Holistic Evaluation of Physical Perception and Reasoning in Large Language Models (Qiu Shi, 2025) View paper
  - [41] Scieval: A multi-level large language model evaluation benchmark for scientific research (Sun, 2024) View paper
  - Medical and Clinical Benchmarks (2 papers)
  - [6] Medxpertqa: Benchmarking expert-level medical reasoning and understanding (Zuo Yuxin, 2025) View paper
  - [44] DiagnosisArena: Benchmarking Diagnostic Reasoning for Large Language Models (Zhu Ya-kun, 2025) View paper
  - Multilingual and Cross-Lingual Evaluation (1 papers)
  - [47] CIF-Bench: A Chinese Instruction-Following Benchmark for Evaluating the Generalizability of Large Language Models (Li Yi-Zhi, 2024) View paper
  - Specialized Reasoning and Knowledge Tasks (2 papers)
  - [34] MORABLES: A Benchmark for Assessing Abstract Moral Reasoning in LLMs with Fables (Marcuzzo, 2025) View paper
  - [50] Time Series Foundation Models: Benchmarking Challenges and Requirements (Meyer, 2025) View paper
- Contamination Impact and Mitigation Strategies
  - Training Data Deduplication and Cleaning (1 papers)
  - [9] Deduplicating training data makes language models better (Katherine Lee, 2022) View paper
  - Synthetic Data and Model Collapse (1 papers)
  - [18] Regurgitative training: The value of real data in training large language models (Zhang Jing-hui, 2024) View paper
  - Reasoning-Driven Synthesis and Mitigation (1 papers)
  - [46] Beyond Memorization: Reasoning-Driven Synthesis as a Mitigation Strategy Against Benchmark Contamination (Terry Jingchen Zhang, 2025) View paper
- Contamination Surveys and Empirical Analyses
  - Survey and Review Papers (5 papers)
  - [5] Benchmark Data Contamination of Large Language Models: A Survey (Xu Cheng, 2024) View paper
  - [10] A survey on data contamination for large language models (Cheng Yu-Xing, 2025) View paper
  - [11] A Comprehensive Survey of Contamination Detection Methods in Large Language Models (Ravaut, 2024) View paper
  - [32] Benchmarking Large Language Models Under Data Contamination: A Survey from Static to Dynamic Evaluation (Simin Chen, 2025) View paper
  - [43] Unveiling the spectrum of data contamination in language models: A survey from detection to remediation (Deng Chun-yuan, 2024) View paper
  - Empirical Contamination Reports (2 papers)

- [4] Generalization or memorization: Data contamination and trustworthy evaluation for large language models (Dong Yihong, 2024) View paper
- [8] An open-source data contamination report for large language models (Yucheng LI, 2024) View paper

- Evaluation Methodology and Benchmark Validity
  - Benchmark Quality and Design Flaws (1 papers)
  - [40] Garbage In, Reasoning Out? Why Benchmark Scores are Unreliable and What to Do About It (Mousavi, 2025) View paper
  - Evaluation Protocol Limitations (2 papers)
  - [15] Benchmark probing: Investigating data leakage in large language models (C Deng, 2023) View paper
  - [38] Training on the benchmark is not all you need (Ni, 2025) View paper

Narrative

Core task: Benchmark contamination detection in large reasoning models. As large language models grow in scale and capability, ensuring that their impressive performance reflects genuine reasoning rather than memorization of benchmark data has become a central concern. The field has organized itself into several complementary branches: Contamination Detection Methods develop techniques to identify whether test data appeared during pretraining, ranging from membership inference approaches like Detecting Pretraining Data[1] to statistical methods such as ConStat[7]. Contamination-Resistant Benchmark Design focuses on creating evaluation sets that remain valid over time, exemplified by continuously updated platforms like LiveCodeBench[2] and dynamic benchmarks. Domain-Specific Contamination Studies examine leakage in specialized areas such as medical reasoning or code generation, while Contamination Surveys and Empirical Analyses provide broad perspectives on the scope and severity of the problem across the ecosystem. Meanwhile, Contamination Evasion and Detection Robustness investigates how models might circumvent detection and how robust current methods truly are.

A particularly active tension exists between detection methods and their limitations: while many studies propose contamination indicators, works like Contamination Detection Limitations[12] and Evading Contamination Detection[16] reveal that sophisticated training procedures can produce contamination-like signals without actual leakage, or conversely, evade existing detection schemes. Fragility Contamination Detection[0] sits squarely within this robustness-focused branch, examining how fragile current detection approaches are when models employ evasion strategies. Its emphasis on adversarial scenarios contrasts with more straightforward detection proposals like Data Contamination Quiz[3], which assumes cooperative evaluation settings. By exploring vulnerabilities in detection pipelines, Fragility Contamination Detection[0] complements the adversarial perspective of Evading Contamination Detection[16], together highlighting that the contamination problem extends beyond simply identifying overlap to understanding the strategic dynamics between model developers and evaluators.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Evading Data Contamination Detection for Language Models is (too) Easy

Authors: MÃ¼ller, Mark Niklas, Jasper Dekoninck, Baader, Maximilian, et al. (13 authors total) | **Year/Venue**: 2024 • arXiv.org | **URL**: View paper

Abstract

Large language models are widespread, with their performance on benchmarks frequently guiding user preferences for one model over another. However, the vast amount of data these models are trained on can inadvertently lead to contamination with public benchmarks, thus compromising performance measurements. While recently developed contamination detection methods try to address this issue, they overlook the possibility of deliberate contamination by malicious model providers aiming to evade detec...

Relationship Analysis

Both papers belong to the 'Evasion Techniques and Vulnerabilities' category, examining how benchmark contamination can be concealed from detection methods in large language models. While the original paper focuses on two specific scenarios in large reasoning models (LRMs)—showing that RL training (particularly GRPO) conceals SFT contamination and that CoT contamination in advanced LRMs evades detection—the candidate paper takes a broader adversarial perspective by proposing EAL, a deliberate contamination technique designed to evade existing detection methods across general LLMs. The key difference is that the original paper analyzes inherent vulnerabilities arising from standard training procedures (SFT+RL pipeline), whereas the candidate paper explicitly develops an evasion attack method from a malicious actor's viewpoint.

Contributions Analysis

Overall novelty summary. The paper investigates how reinforcement learning training can conceal benchmark contamination signals in large reasoning models, focusing on two training stages: supervised fine-tuning and RL optimization. It resides in the 'Evasion Techniques and Vulnerabilities' leaf, which contains only two papers total. This is a notably sparse research direction within the broader taxonomy of 50 papers across 24 leaf nodes, suggesting that adversarial perspectives on contamination detection remain underexplored compared to detection method development or benchmark design.

The taxonomy reveals substantial activity in adjacent areas: the parent category 'Contamination Evasion and Detection Robustness' also includes 'Detection Method Evaluation and Limitations' with two papers examining detection method failures. Meanwhile, sibling branches like 'Contamination Detection Methods' contain 15 papers across multiple detection approaches (black-box statistical methods, performance-based detection, white-box training data analysis). The paper's focus on RL-stage concealment connects to 'Fine-Tuning and RL-Stage Detection' methods but approaches the problem from an adversarial rather than defensive angle, examining how PPO-style objectives enable evasion.

Among 25 candidates examined across three contributions, none were found to clearly refute the paper's claims. The systematic study of contamination across training stages examined 9 candidates with 0 refutations; the RL concealment discovery examined 6 candidates with 0 refutations; and the CoT contamination evasion finding examined 10 candidates with 0 refutations. This suggests that within the limited search scope, the specific focus on RL training as a contamination concealment mechanism and the theoretical analysis of PPO-style importance sampling effects represent relatively unexplored territory, though the search scale precludes definitive conclusions about the broader literature.

The analysis indicates the work addresses a genuine gap in understanding adversarial dynamics between model training and contamination detection, particularly regarding RL optimization phases. However, the limited search scope (25 candidates from semantic search) and the sparse population of the evasion-focused taxonomy leaf mean this assessment reflects top-K semantic matches rather than exhaustive coverage. The novelty appears strongest in mechanistic analysis of how specific RL objectives conceal contamination, though broader claims about detection fragility should be contextualized within the examined candidate set.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Systematic study of benchmark contamination in LRMs across two stages

Description: The authors conduct the first comprehensive investigation of benchmark contamination in Large Reasoning Models, examining two distinct stages: Stage I (pre-LRM) when base models evolve into LRMs via SFT and RL, and Stage II (post-LRM) when contamination with CoT is applied to advanced LRMs as a final step.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Mathador-LM: A Dynamic Benchmark for Mathematical Reasoning on Large Language Models

URL: View paper

Brief Assessment

Mathador-LM[54] addresses benchmark contamination through dynamic instance generation to prevent test-set leakage, rather than studying contamination detection methods across LRM training stages (SFT and RL).

2. DICE: Detecting In-distribution Contamination in LLM's Fine-tuning Phase for Math Reasoning

URL: View paper

Brief Assessment

DICE[45] focuses on in-distribution contamination detection using internal model states for math reasoning tasks, not on systematic investigation of contamination across distinct training stages (pre-LRM and post-LRM) as described in the original contribution.

3. Self-Explore: Enhancing Mathematical Reasoning in Language Models with Fine-grained Rewards

URL: View paper

Brief Assessment

Self-Explore[53] focuses on self-improvement of reasoning capabilities through fine-grained rewards for detecting wrong steps in rationales. It does not address benchmark contamination detection or evaluation integrity in LRMs.

4. Benchmarking Benchmark Leakage in Large Language Models

URL: View paper

Brief Assessment

Benchmarking Benchmark Leakage[52] focuses on detecting benchmark leakage in general LLMs using perplexity and n-gram accuracy metrics, without examining the specific two-stage contamination process (pre-LRM and post-LRM) that characterizes reasoning model development through SFT and RL.

5. Commonsense reasoning with rules, cases, and connectionist models: a paradigmatic comparison

URL: View paper

Brief Assessment

Commonsense Reasoning Paradigms[57] focuses on comparing rule-based, case-based, and connectionist approaches to commonsense reasoning. It does not address benchmark contamination, large reasoning models, or training stage analysis.

6. Pretraining on the test set is no longer all you need: A debate-driven approach to qa benchmarks

URL: View paper

Brief Assessment

Pretraining Test Set[51] focuses on debate-driven evaluation methods to address benchmark contamination in general language models, not on systematic investigation of contamination across distinct training stages (SFT and RL) in Large Reasoning Models specifically.

7. Addressing data challenges in LLM-enhanced software engineering

URL: View paper

Brief Assessment

Data Challenges LLM[55] addresses data leakage in code generation benchmarks (HumanEval) using combinatorial testing, not contamination detection across LRM training stages (SFT and RL). The candidate focuses on preventing benchmark leakage through test design rather than detecting contamination in reasoning model training pipelines.

8. Detecting Data Contamination from Reinforcement Learning Post-training for Large Language Models

URL: View paper

Brief Assessment

Detecting RLHF Contamination[22] focuses specifically on contamination detection during the RL post-training phase using a self-critique method, whereas the original paper examines contamination across two distinct stages (pre-LRM and post-LRM) with emphasis on how RL conceals SFT contamination and how contamination with CoT leaves minimal evidence. The candidate does not address the two-stage framework or the concealment mechanisms that are central to the original contribution.

9. Rethinking the Reasonability of the Test Set for Simultaneous Machine Translation

URL: View paper

Brief Assessment

Rethinking Test Set[56] focuses on simultaneous machine translation evaluation and test set design for translation models, not on benchmark contamination detection in Large Reasoning Models across training stages.

Contribution 2: Discovery that RL training conceals SFT contamination evidence

Description: The authors demonstrate that while SFT contamination is initially detectable, subsequent GRPO training on clean samples conceals contamination evidence. They provide theoretical analysis showing that PPO-style importance sampling and clipping objectives are the root cause of this concealment.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Removing RLHF Protections in GPT-4 via Fine-Tuning

URL: View paper

Brief Assessment

Removing RLHF Protections[73] focuses on removing safety protections via fine-tuning in GPT-4, not on concealing benchmark contamination signals in reasoning models through RL training.

2. Best-of-venom: Attacking rlhf by injecting poisoned preference data

URL: View paper

Brief Assessment

Best-of-venom[70] focuses on poisoning RLHF preference data to manipulate model outputs toward specific entities/sentiments, not on detecting or concealing benchmark contamination from supervised fine-tuning.

3. Training Language Models to Self-Correct via Reinforcement Learning

URL: View paper

Brief Assessment

Self-Correct Reinforcement Learning[68] focuses on training self-correction capabilities in LLMs using online RL, not on detecting or concealing benchmark contamination from prior SFT stages.

4. Teaching Large Language Models to Reason with Reinforcement Learning

URL: View paper

Brief Assessment

Teaching Reasoning RL[69] focuses on comparing RL algorithms for improving reasoning capabilities, not on benchmark contamination detection or concealment effects. The paper does not address contamination signals or detection methods.

5. The Impact of Post-training on Data Contamination

URL: View paper

Brief Assessment

Post-training Contamination Impact[72] examines how post-training (SFT and RL) resurfaces contamination signals that diminish during continued pre-training, whereas the original paper focuses on how RL conceals initially detectable SFT contamination. These are fundamentally different phenomena occurring at different training stages.

6. Reinforcement Learning with Supervised Alignment

URL: View paper

Brief Assessment

Supervised Alignment RL[71] focuses on using reinforcement learning to improve LLM generalization through supervised alignment for reward modeling in question-answering tasks. It does not address benchmark contamination detection or the concealment of SFT contamination signals through RL training.

Contribution 3: Finding that CoT contamination on advanced LRMs evades existing detection methods

Description: The authors reveal that contaminating advanced LRMs with chain-of-thought reasoning in the final training stage yields inflated performance while leaving minimal detectable evidence, causing existing memorization-based detection methods to perform near random guessing across all benchmarks.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. LNE-Blocking: An Efficient Framework for Contamination Mitigation Evaluation on Large Language Models

URL: View paper

Brief Assessment

LNE-Blocking[63] focuses on contamination mitigation evaluation (restoring model performance after contamination) rather than contamination detection methods. The candidate does not address chain-of-thought contamination or detection evasion in advanced reasoning models.

2. Program Verification to Defend Chain-of-Thought Attacks for LLM Services

URL: View paper

Brief Assessment

Program Verification Defense[67] focuses on defending against adversarial attacks on chain-of-thought reasoning (e.g., BadChain attacks), not on detecting benchmark contamination in training data. The two papers address fundamentally different problems in the CoT domain.

3. GUARD: Dual-Agent based Backdoor Defense on Chain-of-Thought in Neural Code Generation

URL: View paper

Brief Assessment

GUARD[62] addresses backdoor attacks in chain-of-thought models for code generation, not benchmark contamination detection in reasoning models. The technical focus and problem domain are fundamentally different.

4. When chain of thought is necessary, language models struggle to evade monitors

URL: View paper

Brief Assessment

CoT Evading Monitors[60] focuses on chain-of-thought monitoring for AI safety (detecting harmful reasoning), not on benchmark contamination detection in language models. These are distinct problems with different objectives.

5. BadThink: Triggered Overthinking Attacks on Chain-of-Thought Reasoning in Large Language Models

URL: View paper

Brief Assessment

BadThink[64] focuses on backdoor attacks that induce overthinking behavior in CoT-enabled LLMs through training-time data poisoning, not on benchmark contamination detection evasion. The technical mechanisms and objectives are fundamentally different from the original paper's contamination detection study.

6. Can large language models detect errors in long chain-of-thought reasoning?

URL: View paper

Brief Assessment

Detecting CoT Errors[58] focuses on evaluating critique abilities to detect errors within long chain-of-thought reasoning processes, not on benchmark contamination detection or how contamination evades detection methods.

7. Chain-of-Thought Hijacking

URL: View paper

Brief Assessment

CoT Hijacking[65] focuses on jailbreaking safety mechanisms through benign reasoning padding, not on benchmark contamination detection evasion. The technical mechanisms and problem domains are fundamentally different.

8. Shadowcot: Cognitive hijacking for stealthy reasoning backdoors in llms

URL: View paper

Brief Assessment

Shadowcot[61] focuses on backdoor attacks that manipulate CoT reasoning chains to produce adversarial outputs, not on benchmark contamination detection evasion. The technical mechanisms and threat models differ fundamentally.

9. Darkmind: Latent chain-of-thought backdoor in customized llms

URL: View paper

Brief Assessment

Darkmind[59] focuses on backdoor attacks that manipulate chain-of-thought reasoning in customized LLMs, not on benchmark contamination detection. The candidate addresses adversarial manipulation of reasoning processes, while the original paper investigates data contamination and its detectability in evaluation benchmarks.

10. Thought Purity: A Defense Framework For Chain-of-Thought Attack

URL: View paper

Brief Assessment

Thought Purity[66] addresses backdoor prompt injection attacks on chain-of-thought reasoning in LRMs, not benchmark data contamination detection. The candidate focuses on defending against malicious triggers that subvert reasoning processes, while the original paper investigates how contaminated training data evades memorization-based detection methods.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] On The Fragility of Benchmark Contamination Detection in Reasoning Models View paper
- [1] Detecting pretraining data from large language models View paper
- [2] LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code View paper
- [3] Data contamination quiz: A tool to detect and estimate contamination in large language models View paper
- [4] Generalization or memorization: Data contamination and trustworthy evaluation for large language models View paper
- [5] Benchmark Data Contamination of Large Language Models: A Survey View paper
- [6] Medxpertqa: Benchmarking expert-level medical reasoning and understanding View paper
- [7] ConStat: Performance-Based Contamination Detection in Large Language Models View paper
- [8] An open-source data contamination report for large language models View paper
- [9] Deduplicating training data makes language models better View paper
- [10] A survey on data contamination for large language models View paper
- [11] A Comprehensive Survey of Contamination Detection Methods in Large Language Models View paper
- [12] Towards data contamination detection for modern large language models: Limitations, inconsistencies, and oracle challenges View paper
- [13] Rethinking Benchmark and Contamination for Language Models with Rephrased Samples View paper
- [14] PaCoST: Paired Confidence Significance Testing for Benchmark Contamination Detection in Large Language Models View paper
- [15] Benchmark probing: Investigating data leakage in large language models View paper
- [16] Evading Data Contamination Detection for Language Models is (too) Easy View paper
- [17] PHYBench: Holistic Evaluation of Physical Perception and Reasoning in Large Language Models View paper
- [18] Regurgitative training: The value of real data in training large language models View paper
- [19] Investigating Data Contamination in Modern Benchmarks for Large Language Models View paper
- [20] AntiLeakBench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge View paper
- [21] Proving Test Set Contamination in Black Box Language Models View paper
- [22] Detecting Data Contamination from Reinforcement Learning Post-training for Large Language Models View paper
- [23] BeyondBench: Benchmark-Free Evaluation of Reasoning in Language Models View paper
- [24] Dynamic Benchmarking of Reasoning Capabilities in Code Large Language Models Under Data Contamination View paper
- [25] How Contaminated Is Your Benchmark? Quantifying Dataset Leakage in Large Language Models with Kernel Divergence View paper
- [26] Truce: Private benchmarking to prevent contamination and improve comparative evaluation of llms View paper
- [27] Is Your Benchmark (Still) Useful? Dynamic Benchmarking for Code Language Models View paper
- [28] Estimating contamination via perplexity: Quantifying memorisation in language model evaluation View paper
- [29] MMLU-CF: A Contamination-free Multi-task Language Understanding Benchmark View paper
- [30] Beyond Correctness: Benchmarking Multi-dimensional Code Generation for Large Language Models View paper
- [31] The Emperor's New Clothes in Benchmarking? A Rigorous Examination of Mitigation Strategies for LLM Benchmark Data Contamination View paper
- [32] Benchmarking Large Language Models Under Data Contamination: A Survey from Static to Dynamic Evaluation View paper
- [33] Putnam-AXIOM: A Functional and Static Benchmark for Measuring Higher Level Mathematical Reasoning in LLMs View paper
- [34] MORABLES: A Benchmark for Assessing Abstract Moral Reasoning in LLMs with Fables View paper
- [35] Simulating Training Data Leakage in Multiple-Choice Benchmarks for LLM Evaluation View paper
- [36] Treeeval: Benchmark-free evaluation of large language models through tree planning View paper

- [37] RealMath: A Continuous Benchmark for Evaluating Language Models on Research-Level Mathematics View paper
- [38] Training on the benchmark is not all you need View paper
- [39] KIEval: A Knowledge-grounded Interactive Evaluation Framework for Large Language Models View paper
- [40] Garbage In, Reasoning Out? Why Benchmark Scores are Unreliable and What to Do About It View paper
- [41] Scieval: A multi-level large language model evaluation benchmark for scientific research View paper
- [42] Training data leakage analysis in language models View paper
- [43] Unveiling the spectrum of data contamination in language models: A survey from detection to remediation View paper
- [44] DiagnosisArena: Benchmarking Diagnostic Reasoning for Large Language Models View paper
- [45] DICE: Detecting In-distribution Contamination in LLM's Fine-tuning Phase for Math Reasoning View paper
- [46] Beyond Memorization: Reasoning-Driven Synthesis as a Mitigation Strategy Against Benchmark Contamination View paper
- [47] CIF-Bench: A Chinese Instruction-Following Benchmark for Evaluating the Generalizability of Large Language Models View paper
- [48] LastingBench: Defend Benchmarks Against Knowledge Leakage View paper
- [49] VAR-MATH: Probing True Mathematical Reasoning in LLMS via Symbolic Multi-Instance Benchmarks View paper
- [50] Time Series Foundation Models: Benchmarking Challenges and Requirements View paper
- [51] Pretraining on the test set is no longer all you need: A debate-driven approach to qa benchmarks View paper
- [52] Benchmarking Benchmark Leakage in Large Language Models View paper
- [53] Self-Explore: Enhancing Mathematical Reasoning in Language Models with Fine-grained Rewards View paper
- [54] Mathador-LM: A Dynamic Benchmark for Mathematical Reasoning on Large Language Models View paper
- [55] Addressing data challenges in LLM-enhanced software engineering View paper
- [56] Rethinking the Reasonability of the Test Set for Simultaneous Machine Translation View paper
- [57] Commonsense reasoning with rules, cases, and connectionist models: a paradigmatic comparison View paper
- [58] Can large language models detect errors in long chain-of-thought reasoning? View paper
- [59] Darkmind: Latent chain-of-thought backdoor in customized llms View paper
- [60] When chain of thought is necessary, language models struggle to evade monitors View paper
- [61] Shadowcot: Cognitive hijacking for stealthy reasoning backdoors in llms View paper
- [62] GUARD: Dual-Agent based Backdoor Defense on Chain-of-Thought in Neural Code Generation View paper
- [63] LNE-Blocking: An Efficient Framework for Contamination Mitigation Evaluation on Large Language Models View paper
- [64] BadThink: Triggered Overthinking Attacks on Chain-of-Thought Reasoning in Large Language Models View paper
- [65] Chain-of-Thought Hijacking View paper
- [66] Thought Purity: A Defense Framework For Chain-of-Thought Attack View paper
- [67] Program Verification to Defend Chain-of-Thought Attacks for LLM Services View paper
- [68] Training Language Models to Self-Correct via Reinforcement Learning View paper
- [69] Teaching Large Language Models to Reason with Reinforcement Learning View paper
- [70] Best-of-venom: Attacking rlhf by injecting poisoned preference data View paper
- [71] Reinforcement Learning with Supervised Alignment View paper
- [72] The Impact of Post-training on Data Contamination View paper
- [73] Removing RLHF Protections in GPT-4 via Fine-Tuning View paper